Assignment 2 Write Up

1) Grid World :- (actions, rewards are as in Assignment 1) which is in the environment

Policy Representation:

i) State-Action Map:- A mapping from each state-action pair to an index in the parameter vector "theta" is created, which allows each state-action pair to have a corresponding weight in the policy parameter vector "theta".

ii) Action Selection:- The action selected given a state are computed using the softmax policy according to its probability.

Hyperparameters:-

i) Learning Rate (α) = 0.005 is selected here for REINFORCE and 0.005 also for Baseline REINFORCE

ii) Discount factor (γ) = 0.09 for both

iii) Baseline Learning Rate (β) = 0.01 is used.

(To stabilize training I tried to normalize rewards)

2) Cart-Pole:- (actions, rewards are as in Assignment 1) which is in the environment

Policy Representation:-

i) State Action Mapping:- A parameter matrix "Theta" with dimensions (state-dim) × (no. of actions). This way each state-action pair has a weight in the policy matrix "Theta".

ii) Action Selection:- $\pi(a|s;\theta) = \frac{\exp(s \cdot \theta_a)}{\sum_b \exp(s \cdot \theta_b)}$

This way we defined the softmax policy. Actions are sampled according to this probability.

Hyperparameters:-

i) Learning Rate (α) = 0.005 in REINFORCE and 0.0001 in Baseline REINFORCE (for avoiding large values that my console can't handle)

ii) Discount Factor (γ) = 0.9 for REINFORCE 0.1 for Baseline REINFORCE because of similar reasons.

iii) Baseline Learning factor (β) = 0.01

(To stabilize training I tried to normalize rewards)